

DOCUMENT RESUME

ED 177 174

TM 009 580

AUTHOR Jensen, Arthur R.
 TITLE The Nature of the Average Difference Between Whites and Blacks on Psychometric Tests: Spearman's Hypothesis.
 PUB DATE [78]
 NOTE 46p.
 EDRS PRICE MF01/PC02 Plus Postage.
 DESCRIPTORS Comparative Testing; Culture Free Tests; *Factor Structure; Intelligence; *Intelligence Differences; *Intelligence Factors; Intelligence Tests; *Racial Differences; Research Problems; Test Interpretation
 IDENTIFIERS *Spearman (Charles)

ABSTRACT

Charles Spearman originally suggested in 1927 that the varying magnitudes of the mean differences between whites and blacks in standardized scores on a variety of mental tests are directly related to the size of the tests' loadings on g, the general factor common to all complex tests of mental ability. Several independent large-scale studies involving factor analysis and the extraction of a g factor from a number of diverse tests given to white and black samples show significant correlations between tests' g loadings and the mean white-black difference (expressed in standard score units) on the tests, thus substantiating Spearman's hypothesis. The average white-black difference on diverse mental tests is interpreted as essentially a difference in Spearman's g, rather than as a difference in the more specific factors peculiar to any particular content, knowledge, acquired skills, or type of test.
 (Author/CTM)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED177174

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

The Nature of the Average Difference Between
Whites and Blacks on Psychometric Tests:
Spearman's Hypothesis

Arthur R. Jensen

University of California, Berkeley

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Arthur R. Jensen

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

ABSTRACT

Charles Spearman originally suggested in 1927 that the varying magnitudes of the mean differences between whites and blacks in standardized scores on a variety of mental tests are directly related to the size of the tests' loadings on *g*, the general factor common to all complex tests of mental ability. Several independent large-scale studies involving factor analysis and the extraction of a *g* factor from a number of diverse tests given to white and black samples show significant correlations between tests' *g* loadings and the mean white-black difference (expressed in standard score units) on the tests, thus substantiating Spearman's hypothesis. The average white-black difference on diverse mental tests is interpreted as essentially a difference in Spearman's *g*, rather than as a difference in the more specific factors peculiar to any particular content, knowledge, acquired skills, or type of test.

Tm009580

**The Nature of the Average Difference Between
Whites and Blacks on Psychometric Tests:
Spearman's Hypothesis**

Arthur R. Jensen

University of California, Berkeley

Representative samples of native-born whites and blacks in the United States generally show an average difference of about one standard deviation between their respective distributions of scores on standard psychometric tests of general intelligence and tests of scholastic aptitude and achievement (Jensen, 1973, Ch.7; Shuey, 1966). One standard deviation difference between the means of two approximately normal distributions with approximately equal SDs corresponds to a median overlap of about 16 percent, i.e., 16 percent of the scores in the lower distribution surpass the median score (or 50th percentile) of the higher distribution.

Not all psychometric tests of ability, however, show the same mean difference (in SD or σ units) or median overlap between the races even when the very same samples are

compared on various tests. There is significant variation in the magnitudes of the mean white-black difference from one test to another (Loehlin, Lindzey & Spuhler, 1975). For example, the racial groups' means seem to differ in varying degrees on tests of various abilities, such as the Verbal, Reasoning, Spatial, and Numerical subscales of the Differential Aptitude Tests (Lesser, Fifer, & Clark, 1965). The groups also show a much smaller average difference on what Jensen (1973b, 1974a) terms Level I abilities (short-term memory and rote learning) than on Level II abilities (abstract reasoning).

Differential psychologists have made little systematic attempt to understand these statistically significant and fairly consistent variations among tests in the size of the mean white-black difference. For many years the most popular explanations have involved cultural and linguistic differences. The tests showing the largest group differences were claimed to be more highly loaded with the cultural content of the white middle-class, and the standard English used in verbal tests was claimed to be a less familiar and less appropriate medium for blacks. The explanatory power of these two hypotheses, however, has utterly failed when the predictions that should logically follow from them have been empirically tested. No study so far has succeeded in showing that when items from standard tests are

classified or rated by expert judges in terms of the items' cultural loadings, independently of any knowledge of the actual item statistics, that the ratings of items' cultural loadings are positively related to the items' discriminability between whites and blacks. In fact, McGurk (1951, 1953a,b) found just the opposite. McGurk asked a panel of 76 judges, including professors of psychology and sociology, to classify 226 items from several well-known standardized tests of general intelligence into one of three categories: least cultural, neutral, most cultural. The 164 items on which there was highest agreement among the judges as to the items' being most or least cultural were administered as a test to large samples of white and black high school pupils. It turned out that the mean white-black difference on the test comprised of items classified as "least cultural" was almost twice as great as the mean white-black difference on the test comprised of items classified as "most cultural." Obviously there must be some property on which these two classes of items differ besides their judged cultural loading that would account for this surprising result. It would seem highly implausible to argue that the blacks were the more culturally advantaged group with respect to the item content of these traditional standard tests.

My conjecture is that when the most and the least culturally loaded test items are roughly matched for difficulty

(i.e., percent failing the item) in either racial group, the least culturally loaded items (e.g., figural analogies) will involve abstract reasoning or problem solving as contrasted with the most culturally loaded items, which will involve the sheer recall of specific acquired knowledge (e.g., information and vocabulary). According to this hypothesis, the size of the white-black difference would be more related to the reasoning dimension than to the cultural loading dimension.

The claim that the style of language used in most standard verbal tests contributes to the racial group difference should lead to the expectation of a greater white-black difference on verbal than on nonverbal tests. However, the massive evidence on this issue unequivocally fails to substantiate this expectation. McGurk (1975) has reviewed virtually the entire published literature between 1951 and 1970 regarding the median overlap between the score distributions of blacks and whites on verbal and nonverbal intelligence tests. The percentage overlap was greater for verbal (19%) than for nonverbal tests (15%) -- a difference significant beyond the .01 level. Thus, in actual fact, blacks perform slightly better, relative to whites, on verbal than on nonverbal tests. Jensen (1974b) found that when items from a quite culture-loaded verbal test (the Peabody Picture Vocabulary Test) and items from a relatively culture-reduced

nonverbal test (Haven's Colored Progressive Matrices) were perfectly matched for difficulty in a white sample of elementary school children, and then were administered to a sample of black children in the same grades, there was no significant difference between the blacks' mean scores on the verbal and nonverbal tests. In other words, verbal and nonverbal tests which are perfectly matched in difficulty at the item level for whites are thereby also matched in difficulty for blacks. Obviously the mean white-black difference in the test scores is not closely linked to the verbal-nonverbal dimension of test characteristics.

If variation in the mean white-black difference on various tests cannot be mainly attributed either to variation in the tests' cultural loading or to the degree of the tests' dependence on language, then we must enquire as to which other characteristic of tests or test items is primarily responsible for the racial discrimination.

Spearman's Hypothesis

Charles Spearman, in 1927, suggested a hypothesis of the essential nature of the white-black difference, which, as far as I can determine, has never been subjected to empirical investigation beyond Spearman's original observation. In commenting on a study that involved ten different mental tests administered to large samples of black and

white American children, Spearman noted that the blacks, on average, obtained lower scores than whites on all ten tests. But Spearman also noted that the mean difference ". . . was most marked in just those [tests] which are known to be most saturated with g" (Spearman, 1927, p.379). Since Spearman's observation was presumably based on a rather limited and unreplicated set of data, it seems best to refer to it, not as an empirical generalization, but as a hypothesis. I shall henceforth refer to it as Spearman's hypothesis.

Spearman's hypothesis that magnitudes of white-black mean differences on various mental tests are directly related to the tests' g loadings, if substantiated by subsequent studies, would be an important and unifying discovery in the study of race differences in mental abilities. The Spearman hypothesis, if true, would mean that the white-black difference in test scores is not essentially attributable to idiosyncratic cultural or linguistic peculiarities in this or that test, but to a general factor that all mental tests measure in common, but which some tests measure to a greater degree than others.

No data, so far, have been collected specifically for the purpose of testing Spearman's hypothesis. But I have searched the psychometric literature for relevant data and have found several studies containing appropriate data which

may be analyzed to determine if the results are predominantly consistent or inconsistent with Spearman's hypothesis.

Methodological Desiderata

An obvious test of Spearman's hypothesis would be the demonstration of a significant relationship or correlation between (a) the *g* factor loadings of various tests and (b) the mean white-black differences (in standardized units) on the various tests.

The first principal component would seem to be the least contentious general factor that we can extract from the matrix of intercorrelations among a number of tests (or test items). The tests' loadings on the first principal component are, in fact, the weighting factors which will maximize the total variance of a linear composite score comprised of the sum of the standardized scores on each of the various tests. In practice we find that the first principal component usually differs negligibly from the first principal factor extracted in a common factor analysis. Also if the test battery is fairly diverse in the types of mental tests included and no one type is over-represented, the first principal component of the battery is highly similar to the second-order *g* factor derived from a hierarchical factor analysis, i.e., the first factor in a common factor

analysis of the intercorrelations among the obliquely rotated first-order factors. Therefore, in all of the following analyses I shall regard the first principal component as the general or g factor of the given battery of tests, and for the sake of simplicity I shall hereafter refer to the tests' loadings on the first principal component as the tests' g loadings.

It should be understood, of course, that the first principal component of any given collection of mental tests (or other measurements) does not necessarily represent the same general factor as Spearman's g , or the same general factor that would be extracted from some quite different collection of tests. However, it turns out, in fact, that different batteries of tests, if they comprise a considerable number and diversity of tests, yield up highly similar g factors. That is to say, the g factor scores derived from the different test batteries are highly correlated with one another. Moreover, examination of the nature of the tests showing the highest g loadings in any battery usually reveals that the items in these most g loaded tests formally reflect Spearman's characterization of g as the capability for abstract reasoning, that is, "the eduction of relations and correlates," to quote Spearman's famous phrase. The inferential ability reflected in highly g loaded test items has operated either largely in the person's past (as in the

acquisition of vocabulary and general information), prior to taking the test, or largely in the immediate test situation itself (as in solving novel figure analogies or progressive matrices). Cattell (1963) has characterized these two aspects of g as crystallized and fluid, or g_c and g_f , respectively.

The average difference between two groups on a given test must, of course, be expressed in standardized units, if it is to be meaningfully compared with the average group difference on some other test. I have used as the standardized unit the square root of the variance within groups, which I shall henceforth refer to as a sigma (σ) unit. This σ unit is equivalent to the weighted average standard deviation within groups, the weights being the respective sizes of two samples, i.e., the sigma unit for two groups, A and B, is

$$\sigma = [(N_A \sigma_A^2 + N_B \sigma_B^2) / (N_A + N_B)]^{1/2},$$

where σ_A and σ_B are the standard deviations of groups A and B, respectively, and N_A and N_B are the numbers of persons in each group. The mean difference between the groups expressed in σ units is simply $\bar{d}_\sigma = (\bar{X}_A - \bar{X}_B) / \sigma$.

One kind of evidence supporting the Spearman hypothesis, then, would consist of a significant coefficient of correlation (or other form of relationship) between

tests' g loadings and the mean white-black differences ($\bar{d}_g$) on the tests.

Four caveats should be observed in the application of this method.

1. The first principal component or g factor in a battery of tests (or items) should be essentially the same factor, within the limits of sampling error, in both racial groups. This can be examined in two ways:

First, we can determine the degree of similarity between groups in the pattern of factor loadings over the various tests by obtaining the Pearson r between the two sets of loadings. A high and significant correlation means the relative magnitudes of the factor loadings on the various tests are similar for both racial groups. A possible problem with this arises if all of the tests are nearly equally loaded on g . Then the slight differences in g loadings will not form a sufficiently reliable pattern to allow a substantial correlation between groups.

Second, we can perform a chi square test of the overall significance of the difference between the two sets of factor loadings, a test which is simultaneously sensitive to a group difference in the pattern of loadings as well as to the absolute magnitudes of the loadings. If the null

hypothesis is not rejected by this chi square test, the g factor is presumed to be the same in both groups. The chi square test in this case is:

$$\chi^2 = \left(\sum_1^n (\underline{z}_A - \underline{z}_B)^2 \right) / \left(\frac{n}{N_A - 3} + \frac{n}{N_B - 3} \right),$$

where χ^2 has 1 degree of freedom, and $\underline{z}$ is Fisher's z transformation of the Pearson correlation coefficient (in this case the factor loading), n is the number of variables, and N_A and N_B are the sample sizes of groups A and B. If the obtained χ^2 with 1 df is less than 3.84, which is the value of χ^2 required for the .05 level of significance, we may accept the null hypothesis that the tests' loadings on the g factor do not differ between the two racial groups.

2. The racial samples being compared should not have been selected in terms of any highly g loaded criterion. For example, we could not properly test the Spearman hypothesis by using white and black students in a highly selective college which applies the same selection criteria to all applicants regardless of race, since such selection for academic aptitude would tend to equalize the racial group means on the most g loaded tests. Moreover, the g factor extracted from tests given to such a highly selected group would be considerably diminished, and probably distorted, as compared with the g extracted from the same tests

given to random samples of either racial population.

3. Test reliability affects both factor loadings and group mean differences (in σ units). Both variables are attenuated (i.e., decreased) by measurement error. Hence, if reliability differs markedly from one test to another in the battery of tests, the correlation between the tests' g loadings and the mean racial difference on the tests will be suprisingly inflated by the common influence of unreliability (measurement error) on both variables. This is probably not a serious drawback if all of the tests have quite similar reliabilities or if there is no systematic relationship between tests' reliabilities and tests' intrinsic g loadings (i.e., the g loadings after correction for attenuation). Of course, it is always most desirable, when the test reliabilities are known, to correct both the g loadings and the σ differences for attenuation. This is simply accomplished by dividing each variable (i.e., the g loading and the σ difference) by the square root of the test's reliability coefficient.

4. Some caution must be exercised in the theoretical interpretation of a high correlation between tests' g loadings on the first principal component and the mean σ difference between groups. Not every collection of tests necessarily yields a first principal component which can be prop-

erly interpreted as Spearman's g in the psychological sense intended by Spearman. The first principal component is affected by variation in psychometric sampling; different collections of tests will result in somewhat different first principal components, especially if each collection has a concentration of highly similar tests that differ quite markedly from the tests in other collections. Thus one must look for evidence that the first principal component can reasonably be interpreted as Spearman's g . Marker tests with known high g saturations in many other factor analytic studies may serve as an important indicator. Another potent indicator is the degree of relationship between various tests' g loadings and these same tests' correlations with IQ or total scores on the best available tests of general intelligence. It seems safe to say that most of the variance (probably as much as 75 to 85 percent) in total scores on standard omnibus intelligence tests represents Spearman's g . Therefore, the loadings on the first principal component of a collection of cognitive tests should be quite highly related to the correlations of these tests with total scores on tests of general intelligence or IQ if the first principal component is to be properly interpreted as Spearman's g . Finally, of course, we should inquire as to the nature of the two or three tests that show the highest loadings on the first principal component of our collection of tests to see

if these highly loaded tests display the properties of relation education, abstractness, or transformational complexity which psychologically best characterize Spearman's *g*.

If the psychological interpretation of the first principal component (as contrasted with its purely mathematical interpretation) is in doubt, then what would be the meaning of a high degree of relationship between the factor loadings (derived within either white or black samples) of the various tests and the sizes of the white-black mean differences on those tests? Such a relationship could, of course, neither confirm nor disconfirm Spearman's hypothesis, if there is doubt that the first principal component is actually Spearman's *g*. However, such a finding would mean, at the very least, that whatever linear composite of these various tests discriminates the most among individuals within each racial group also discriminates the most between the means of the two racial groups. This implies, of course, that individual differences within the racial groups and the mean difference between the racial groups are factorially the same (or highly similar), whatever the psychological nature of the factor may be. In other words, the first principal component of this battery of tests discriminates between white and black persons on the same basis as it discriminates between persons of the same race, whether or not the first principal component is psychologically interpretable

as Spearman's g . This would be the expected outcome, of course, if the tests in the battery were not racially biased in discriminating individual differences.

Factor Analytic Studies of White and Black Samples

Several recent studies using a variety of tests in quite large samples of whites and blacks, unselected for general intelligence, are here examined with respect to Spearman's hypothesis that there exists a positive relationship between a test's g loading and the magnitude of the mean white-black difference in test scores.

Nichols (1972). A doctoral study by Nichols (Note 2) provides relevant data on seven subscales of the Wechsler Intelligence Scale for Children (WISC) in addition to six other cognitive ability and achievement tests given to large samples of white ($N=1940$) and black ($N=1460$) seven-year old children in several large cities in the United States. All of the subjects were participants in a large-scale longitudinal study conducted by the National Institutes of Health. The subjects were enlisted in 12 public hospitals at the time of their mothers' pregnancy, and are a fairly representative sample of the populations served by these large city hospitals, a population which Nichols describes as "skewed somewhat to the lower end" in social class.

Nichols notes (p.83) that the intercorrelations among the 13 tests are highly similar in the white and black samples, as indicated by a correlation of +0.954 between the two correlation matrices. Obviously the factor structure of these variables is bound to be highly similar for whites and blacks, and factor analysis bears this out. The g factor loadings extracted from these correlation matrices are shown in Table 1 along with the mean white black differences

 Insert Table 1 about here

expressed in σ units. A chi square test of the significance of the overall white-black differences in g loadings of the 13 tests yields a $\chi^2=1.31$, which is nonsignificant even with these very large samples. The Pearsonian correlation between the white and black g loadings is +0.98, which for 12 df is significant beyond the .001 level.

The correlation between the mean white-black differences and the g loadings on the 13 tests is +0.69 ($p<.01$) for whites and +0.71 ($p<.01$) for blacks. If we compare (a) the mean white-black difference (in σ units) on the six tests with the highest g loadings with (b) the mean white-black difference on the seven tests with the lowest g loadings, we have $0.70\sigma-0.46\sigma=0.24\sigma$, which is a highly significant difference (noncorrelated $t=6.92$, $df=3398$, $p<.001$).

Table 1

Mean Difference (in Sigma Units) Between Seven-Year Old Whites ($N = 1940$) and Blacks ($N = 1460$), and Tests' g Loadings in Each Sample

Test	Mean W-B Difference ^{a/}	Loading on g	
		White	Black
1. Bender-Gestalt	.69	.60	.61
2. WISC Information	.37	.67	.69
3. WISC Comprehension	.41	.47	.49
4. WISC Vocabulary	.85	.67	.71
5. WISC Digit Span	.45	.65	.63
6. WISC Picture Arrangement	.71	.64	.64
7. WISC Block Design	.66	.55	.58
8. WISC Coding	.17	.37	.31
9. I.T.P.A. ^{b/}	.96	.75	.75
10. Draw-a-Man	.11	.49	.52
11. WRAT Spelling ^{c/}	.73	.79	.78
12. WRAT Reading	.73	.78	.78
13. WRAT Arithmetic	.55	.79	.77
Percent of Total Variance Accounted for		41.6	42.1

^{a/} Mean White-Black Difference in σ units. The difference is

$$\sigma = \frac{(\bar{X}_W - \bar{X}_B) \rightarrow}{\left[\frac{(N_W \sigma_W^2 + N_B \sigma_B^2)}{(N_W + N_B)} \right]^{1/2}}$$

^{b/} Illinois Test of Psycholinguistic Abilities (Auditory-Vocal Association)

^{c/} Wide Range Achievement Tests

Source: Secondary analyses of data from Nichols, 1972, Tables 18, 19, 20.

Thus Spearman's hypothesis would seem to be borne out by these data. Certainly it is not contradicted. Yet there is a question, since the *g* loadings of some of these tests are not in close agreement with their *g* loadings where they have been factor analyzed in other contexts. The present *g* factor appears to be overly weighted on verbal and educational achievement variables and rather closely resembles what British factor analysts have termed the *V:ed* (verbal-educational) factor. We are safest in concluding only that the white-black differences are quite substantially correlated with the tests' *g* loadings in this particular battery, for both whites and blacks. This *g* is undoubtedly highly correlated with, but not perfectly representative of, Spearman's *g*. At the very least, however, the results indicate that those tests which best discriminate individual differences among whites are the same tests that best discriminate individual differences among blacks, and are also the same tests that discriminate the most between whites and blacks.

Hennesy and Merrifield (1976). Nearly 3000 high school seniors desiring to enter an open-admission community college in New York City were given the College Entrance Examination Board Comparative Guidance and Placement Program battery, which consists of ten subtests designed to measure (1) vocabulary, (2) reading comprehension, (3) grammatical

rules and usage, (4) quantitative reasoning, (5) following complex directions, (6) perceptual speed and accuracy, (7) inductive reasoning, (8) short-term memory, (9) spatial reasoning, and (10) information and concepts from broadly defined technical areas. All the tests were designed to "minimize the effects of formal classroom preparation and to emphasize material that should be familiar to all high school graduates" (Hennessey & Merrifield, 1976, p.755).

Four ethnic groups are represented in this sample of high school seniors: 431 black, 163 Hispanic, 573 Jewish, and 1818 Caucasian-Gentile, representing a wide range of socioeconomic status.

Since Hennessey and Merrifield were primarily concerned with ethnic differences in the factor structure of abilities, they partialled socioeconomic status (SES) out of the correlations among all of the ability measures. The SES index was based on family income, the occupation of the main wage earner, and the educational level of both parents. (When SES was included in the factor analysis, it showed no loadings above 0.20 on any of the three ability factors that emerged. This small effect of SES, however, was statistically removed from the factor analyses discussed below.)

The same factor structure emerged in all ethnic groups. According to the authors, "The overall conclusion from this

study is that there are no meaningful differences in the factor structures for the four groups... [w]hile there may be differences in the level of performance on the various tests between different ethnic groups, they seem to be measuring the same abilities across these groups" (p.759). The g factor loadings on the ten tests are correlated +0.91 across the white (Caucasian-Gentile) and black groups, which indicates a high degree of similarity in the pattern of g loadings in these two racial groups. However, a χ^2 test of the overall difference in g loadings is significant ($\chi^2=5.06$, $p<.05$), since g accounts for 74.3% of the total variance in the white sample and 81.8% in the black.

The Spearman hypothesis can be tested on these data, using the SES-adjusted means (i.e., SES is statistically held constant for whites [Caucasian-Gentiles] and blacks) provided in Hennessy's doctoral dissertation (Note 1). I have correlated these mean white-black differences (in σ units) with the loadings on the g factor extracted from the pooled ethnic groups. The correlation is +0.74 ($p<.01$). The mean white-black difference on the five tests with the highest g loadings is 0.77σ as contrasted with the white-black difference of 0.52σ on the five tests with the lowest g loadings -- a difference of 0.25σ , which is significant beyond the .001 level, thus clearly bearing out Spearman's hypothesis. There can be little doubt that the g factor in

this battery comes very close to Spearman's g , since the reasoning tests have the highest loadings. Hennessy (1974) computed factor scores (with SES partialled out) for all subjects on the rotated factors. The three factors, labeled Verbal (likened to g_c), Reasoning (likened to g_f), and Spatial-Technical Information show mean white-black differences of 0.59σ , 0.92σ , and 0.33σ , respectively. This suggests that the white-black difference is greater on the fluid (g_f) than on the crystallized (g_c) aspect of Spearman's g .

Osborne. R. Travis Osborne has provided me with data relevant to fluid and crystallized g . Twelve quite diverse tests were given to white ($N=608$) and black ($N=246$) urban school children in Georgia. Eight of the tests are from the Educational Testing Service's "Kit of Reference Tests for Cognitive Factors" (French, et al., 1963). The tests fall into two categories similar to what Cattell (1971) has characterized as "fluid" and "crystallized" intelligence, or g_f and g_c . The two categories of tests are listed below.

"Fluid""Crystallized"

Cube Comparisons

Calendar Test

Identical Pictures

Arithmetic

Formboard

Vocabulary (Wide Range)

Surface Development

Vocabulary (Heim)

Spatial	Spelling
Paper Folding	
Object Aperture	

The fluid tests are all nonverbal and nonscholastic and do not call upon any specific knowledge acquired outside the testing situation.

This battery of 12 tests was factor analyzed separately in the white and black samples. Chronological age (in months) was partialled out of all of the intercorrelations. The rotated factors clearly divide up into g_f and g_c , which together account for 48.6% of the total variance in whites and 48.5% in blacks. The g_f accounts for 24.7% and 25.1% for whites and blacks, respectively; the corresponding figures for g_c are 23.9% and 23.4%. The loadings on g_f are correlated +0.81 ($p < .01$) between whites and blacks, and the loadings on g_c are correlated +0.93 ($p < .01$). Even with this high degree of similarity between the racial groups on g_f and g_c , a chi square test shows the overall differences in factor loadings between the groups to be significantly different: for g_f , $\chi^2 = 7.76$, $p < .01$; and for g_c , $\chi^2 = 4.36$, $p < .05$. Of course quite small differences in factor loadings, as these are, can be statistically significant with such a large sample ($N = 854$).

The Spearman hypothesis was tested on these data for

both $\underline{g}_f$ and $\underline{g}_c$. The correlation between tests' $\underline{g}_c$ loadings and the mean white-black differences is -0.24 for white $\underline{g}_c$ loadings and -0.02 for the black; both r s are nonsignificant. The correlation between loadings on $\underline{g}_f$ and the white-black difference is +0.56 ($p < .05$) for whites and +0.42 ($p < .10$) for blacks. Thus, the mean white-black differences on these twelve tests are more highly related to the tests' loadings on $\underline{g}_f$ than on $\underline{g}_c$. This result contradicts the popular belief that the white-black difference on tests largely involves differences in past learning as characterized by the "crystallized" component of variance in test scores. However, this study does not lend strong support to the Spearman hypothesis per se, since the correlations are quite low. Either the hypothesis is wrong or the $\underline{g}_f$ and $\underline{g}_c$ extracted from this battery are not very close to Spearman's g . The "crystallized" tests indeed seem quite limited in types of content, and the "fluid" tests seem quite loaded on spatial visualization.

Veroff, McClelland, and Marquis (1971). White and black adults between 18 and 49 years of age were randomly selected in about equal numbers by race and sex from a probability sample of 1,027 households within the city of Detroit, sampled so as to yield a cross-section of each race. Seven ability measures, administered to approximately half of each race sample by either a black or a white

interviewer, were obtained on 179 whites and 186 blacks. Complete intercorrelations among the ability measures for both racial groups are presented for only six of the seven tests: Lorge-Thorndike Sentence Completion, Raven's Standard Progressive Matrices, Ammons Quick Test, and the Digit Symbol (Coding) and Information subtests of the Wechsler Adult Intelligence Scale.

These six tests were factor analyzed separately in the white and black samples. The percentage of total variance accounted for by g is 53.4% for whites and 49.8% for blacks, a nonsignificant difference. The correlation between the whites' and blacks' g loadings on the six tests is +0.82 ($p < .025$). The overall difference between white and black g loadings is nonsignificant ($\chi^2 = 1.37$).

We can check the Spearman hypothesis with the available information. Veroff et al. provide an analysis of variance on each test showing the F ratio for the race main effect; the rank order of these six F ratios is, of course, exactly the same as the rank order of the six tests' mean white-black differences expressed in σ units. So we may test the Spearman hypothesis by performing a rank order correlation p between the tests' g loadings and the F ratios indicating the magnitudes of the white-black differences on the tests. The p is +0.66 for whites and +0.60 for blacks. Since with

only 5 degrees of freedom p must be 0.83 or greater to be significant at the 5% level, in this case a statistically more powerful test of the Spearman hypothesis would be to compare the mean of the race effect F ratios for the three tests having the highest g loadings with the three tests having the lowest g loadings and then perform a significance test on this difference. This is properly done by first transforming the F values to unit-normal z values (see Paulson, 1942) and determining the significance of the difference between the average z s of the three highest and the three lowest g loaded tests. It turns out that this difference is highly significant ($p < .001$) for both whites and blacks, a result which is consistent with the Spearman hypothesis.

Jensen (1977). The 50-item Wonderlic Personnel Test (WPT) was obtained on 204 white and 204 black young adult job applicants for entry level jobs in a company in New York City; no selection was made on age, education, or sex. The white and black means are close to the national norms on the WPT and differ by 0.89 σ . The WPT is a test of general intelligence composed of items involving verbal, nonverbal, and logical reasoning, and factual information. A principal components analysis was done separately on the matrix of 49 item intercorrelations in each random half of each racial group. (Item No.50 was dropped, since there was no variance

on this item in either racial group.) Thus the correlation between the groups' factor loadings could be determined both within and between the racial groups. The correlation within groups affords a baseline for assessing the correlation between groups. The correlation of the 49 *g* factor loadings across the random halves of the white sample is +0.69, and of the black sample it is +0.73. The correlation of *g* loadings across the total white and black samples is +0.68 ($p < .01$). Thus the correlation of *g* loadings between races is about the same as within races.

A test was conducted of the hypothesis that the degree to which the items discriminated between whites and blacks is correlated with the items' *g* loadings. For each racial group, the item *p* values (percent passing) were transformed to normalized *z* values, thereby ordering item difficulty on an interval scale (Guilford, 1954, pp.418-419). The difference between the white and black *z* scores is an index of the items discrimination between the two groups. This item index of racial discrimination is correlated with the items' *g* loadings +0.47 for whites and +0.62 for blacks (both *rs* significant beyond the .01 level). These are quite high correlations when viewed against the correlation of the *z* scores between random halves of the same racial group, which is +0.49 for both whites and blacks. Thus there is a substantial relationship between the size of the items'

loadings on the general factor common to all items in the Wonderlic and the magnitudes of the white-black differences on the items, and this is true whether the g factor is determined in the white or in the black samples. The items that best measure individual differences within each racial group are the same items, by and large, that discriminate the most between the racial groups. There can be little doubt that the first principal component of the WPT is very similar to, if not exactly the same as, Spearman's g .

Factor Analysis of Tests Within and Between Families in White and Black Populations

My research assistants have administered seven different mental tests to practically all of the children between the ages of 6 and 12 years in 744 white and 414 black families in a California school district. The tests are Short-term Memory, Figure Copying, Pictorial Intelligence Test (Lorge-Thorndike Primary), Lorge-Thorndike Verbal IQ and Nonverbal IQ, Vocabulary and Reading Comprehension (Stanford Achievement Tests). I would prefer a larger and more diverse sample of psychometric variables for the comparison of factor structures across different ethnic groups, but these data were not collected originally with that purpose in mind. Yet the data will serve as an illustration of a method for testing an interesting hypothesis, namely, that

the factorial structure of the tests is the same for children sharing the same family background as for children reared in different families. Any test score differences among children in the same family cannot be regarded as being due to differences in cultural background, while such differences between children from different families may or may not reflect cultural differences. Rejection of the hypothesis that test score differences between siblings reared together in the same family are factorially the same as differences between unrelated children reared in different families would, of course, favor the hypothesis that the tests are culture biased.

The differences within families can be factor analyzed by (1) obtaining for each test the difference between the age-standardized scores of every pair of siblings, (2) intercorrelating the sibling differences on the various tests, and (3) factor analyzing the correlation matrix. Such a within-family correlation matrix can contain no variance attributable to differences between families, and consequently no variance due to socioeconomic or cultural factors, as these terms are ordinarily understood. Sibling differences can reflect only genetic variance and the microenvironmental variance among members of the same family. The differences between families can be factor analyzed by (1) obtaining for each test the mean of the

age-standardized scores of all siblings in each family, then (2) intercorrelating the family means on the various tests, and (3) factor analyzing the correlation matrix.

The percentage of the total variance accounted for by the first principal component or *g* factor in each of the four matrices is:

	<u>White</u>	<u>Black</u>
<u>Between Families</u>	66.6%	61.0%
<u>Within Families</u>	51.9%	52.4%

The Between Families *g* accounts for a larger percentage of the total variance because the scores entering into the correlations are more reliable, being the means of two or more siblings on a given test. (The means of any two or more correlated measurements are always more reliable than the single measurements.) By the same token, the Within Family correlations, based on the differences between siblings on a given test, are necessarily less reliable. (Differences between any two imperfectly reliable measurements are always less reliable than either of the single measurements.)

The pattern of *g* loadings on the seven tests is highly similar across the four groups, as shown by the cross-group correlations between the tests' *g* loadings:

	2) W w-f	3) B b-f	4) B w-f
1) White between-families	.95	.84	.97
2) White within-families		.95	.94
3) Black between-families			.88
4) Black within-families			

The above correlations are statistically homogeneous, that is, they do not differ significantly from one another. Thus it appears that the g loadings of these seven tests show a very similar pattern regardless of whether they were extracted from the within-family correlations (which completely excludes cultural and socioeconomic effects in the factor analyzed variance) or from the between-families correlations, for either whites or blacks. The results are consistent with the hypothesis that the intelligence and achievement test score differences between children from different families, whether white or black, involve the same g factor as the test score differences between siblings reared together in the same family. This outcome would seem unlikely if the largest source of variance in these tests, reflected by their g loadings, were strongly influenced by whatever cultural differences that might exist between families, and between whites and blacks.

The Spearman hypothesis can be tested on these data. Since the pattern of g loadings on the seven tests is so

highly similar in both racial groups and for both the between- and within-family conditions, I have averaged the four sets of g loadings (via Fisher's Z transformation) on each test and correlated these seven averages with the mean white-black differences (in σ units) on the seven tests. The correlation is $+0.78$ ($p < .05$). Also, the mean white-black difference on the three most g loaded tests is significantly ($p < .001$) greater than on the three least g loaded tests, consistent with Spearman's hypothesis.

A More Definitive Test of the Spearman Hypothesis

None of the studies reviewed previously provides the kind of precise estimates of test reliability based on the samples in the study (or highly comparable samples) that are needed if we are properly to correct the g loadings and mean differences (in σ units) for attenuation. In most cases, however, it seems unlikely that the lack of correction for attenuation could seriously distort the results, especially where standardized tests are involved. Most such tests have reliability coefficients of close to 0.90, and there is little variation among the reliabilities. Since the correction for attenuation is based on the square root of the reliability coefficient, it would have relatively little effect under these conditions. Moreover, there is no good reason to suppose that tests' reliabilities are systematically

related to the tests' intrinsic *g* loadings across the various studies. Thus it seems most improbable that repeated confirmation of the Spearman hypothesis from one study to another, each based on a different collection of tests, could be an artifact ascribed to variations among the tests' reliabilities. Yet the lack of correction for attenuation may create some room for doubt.

An impressive set of data which permits appropriate corrections for attenuation is provided by the United States Employment Service in its manual and many technical reports on the General Aptitude Test Battery or GATB (U.S. Department of Labor, 1970). This is as nearly an ideal set of data for testing the Spearman hypothesis as can be found at present.

The GATB is a quite comprehensive battery of tests, originally derived by means of factor analytic studies, which yields scores on nine aptitudes that have been found useful, in different weighted combinations, for employment selection for a wide variety of occupations. The aptitudes measured by the GATB are as follows:

G - General Intelligence

V - Verbal Aptitude

N - Numerical Aptitude

- S - Spatial Visualization Aptitude
- P - Form Perception
- Q - Clerical Aptitude
- K - Motor Coordination
- F - Finger Dexterity
- M - Manual Dexterity

The first principal component or g of this battery can be obtained for each of five correlation matrices (i.e., intercorrelations among the nine GATB aptitudes) given in the GATB Manual (Tables 6-5 through 6-9, pp.32-34). The intercorrelations are based on large samples (total $N=27,365$) of employed workers, high school seniors, college freshmen, basic airmen, and applicants, apprentices, and trainees in various jobs. The g loadings of the nine aptitudes in these five samples are so highly similar as to justify averaging them over the five samples. Using analysis of variance, the profile of these averaged g factor loadings on the nine aptitudes is shown to have a profile reliability of 0.97. Thus the nine-point profile or pattern of average g loadings is extremely stable, with very little sampling error.

Another way that we can estimate the extent to which each of the nine aptitudes is related to general intelligence is to determine the correlation of each one with some

standard test devised to measure IQ or general intelligence. Fortunately, the GATB Manual gives the correlations in large adult samples between each of the nine aptitude scores and total IQ (or some equivalent score) on each of twelve well-known standard tests of IQ or general intelligence. Such tests are presumably largely measures of Spearman's g . (The 12 tests are: ACE Psychological Examination, California Test of Mental Maturity, Cattell Culture-Fair Test of g , Raven's Colored Progressive Matrices, the Reasoning Test of the Differential Aptitude Test battery, Henmon-Nelson IQ, Lorge-Thondike IQ, Otis IQ, Beta, School and College Aptitude Test, Wechsler Adult Intelligence Scale, Wonderlic Personnel Test.) I averaged the 12 correlations (via Fisher's Z transformation) between each of the nine GATB aptitudes and the 12 IQ tests. Analysis of variance shows the profile reliability of this nine-point profile of averaged correlations is 0.97.

Next, since the g loadings of the aptitudes as well as their correlations with the IQ tests are affected to some degree by the differing reliabilities (ranging from 0.67 to 0.87) of the nine aptitude variables, we must subject the g loadings and correlations to correction for attenuation, using the average reliabilities of the nine aptitude tests provided in the GATB Manual (p.255). (The nine-point profile of these reliability coefficients has a profile

reliability of 0.94.) It so happens that the least g loaded tests also have the lowest reliabilities, and therefore correction for attenuation is essential to rule out the possibility that the differences between the aptitudes' g loadings or their correlations with the IQ tests are not merely the result of their differing reliabilities. So from here on I shall be dealing only with g loadings and correlations which have been corrected for attenuation.

If the g of the nine aptitudes as extracted by a principal components analysis (the g being the first principal component) is essentially the same as Spearman's g , which we assume is measured by standard tests of IQ, we should expect a considerable degree of resemblance between the profile of g loadings on each of the nine aptitudes and the profile of average correlations with IQ for each of the nine aptitudes. In fact, there is considerable resemblance between these two profiles, the correlation between them being +0.95. These profiles, shown in Figure 1, are obviously very similar.

Insert Figure 1 about here

Finally, in order to test the Spearman hypothesis, we need to determine the mean white-black difference on each of the GATB aptitudes. The needed data were obtained from 24 technical reports provided by the U.S. Employment Service.

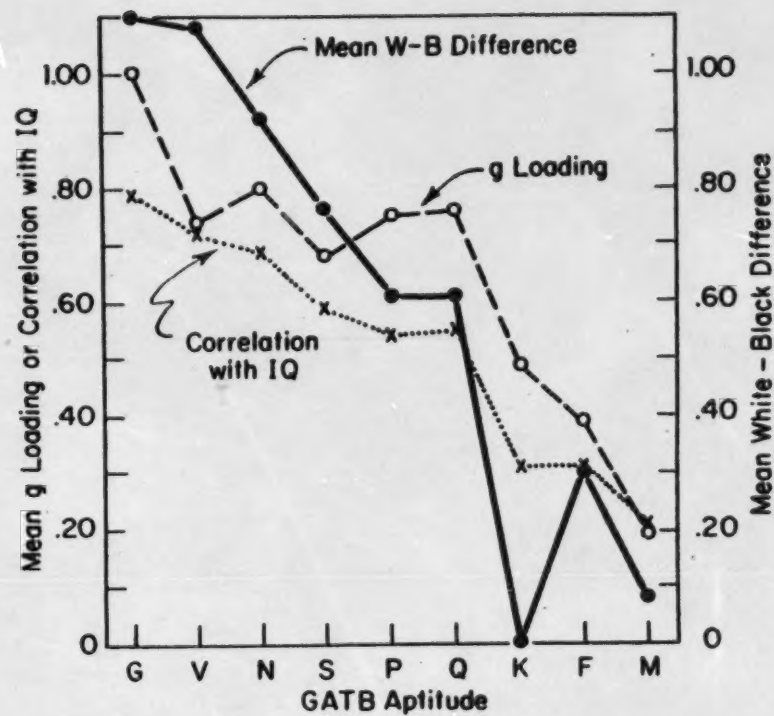


Figure 1. Profiles of (a) the correlations of each of the nine GATB aptitudes with IQ, (b) the *g* factor loadings of each aptitude, and (c) the mean white-black difference (in σ units) on each of the aptitudes.

(See Appendix). These reports give the means and standard deviations on each of the nine GATB aptitudes for large groups of whites and blacks in 24 different occupational categories, comprising a total of 2631 whites and 1521 blacks. (It is noteworthy that studies of the validity of the GATB for predicting job performance criteria do not find a significant difference between the GATB validity coefficients for white and black samples in any of these 24 occupations.) In each occupational group the mean white-black difference was divided by the average within groups standard deviation, so that the mean white-black difference is expressed in σ units. These differences then were averaged (with unit weights) over the 24 occupational groups for each of the nine GATB aptitudes. Since σ differences are also attenuated by unreliability, just as are g loadings and correlations, the white-black σ differences were corrected for attenuation by dividing each σ difference by the square root of the particular aptitude test's reliability. The nine-point profile of these average white-black σ differences (averaged over the 24 occupations) has a profile reliability of 0.99. In other words, the pattern or profile of mean white-black differences across the nine GATB aptitudes is extremely stable and representative of all of the occupational samples.

This profile of white-black differences on the nine

aptitudes is shown in Figure 1. Note its marked similarity to the profiles of g loadings and IQ correlations. An index of the degree of similarity among the three profiles is the correlation (Pearson r) between pairs of profiles:

g loading x correlation with IQ:	$r = 0.95(0.98)$
g loading x white-black difference:	$r = 0.88(0.90)$
Correlation with IQ x white-black difference:	$r = 0.98(1.00)$

The profile intercorrelations after correction for attenuation (using the profile reliabilities) are shown above in parentheses. The one aptitude which shows the largest discrepancy between the otherwise highly similar profiles of g loadings and mean white-black differences, as shown in Figure 1, is Aptitude K: Motor Coordination. Although its g loading in this battery is almost +0.50, there is zero difference between the white and black means. This would happen only if blacks, on average, were superior to whites on the non- g factor(s) measured by the motor coordination test.

These results in general strongly bear out Spearman's hypothesis that the magnitude of white-black differences on various ability tests is directly related to the test's g loading. Hence, achieving a scientific understanding of the nature of the white-black difference in cognitive abilities will depend, in part, upon understanding the nature of

Spearman's g , which is surely one of the most interesting and enduring constructs in all of psychology. Our present knowledge about the nature of g is quite limited to descriptions of the types of tests or problems that are most g loaded and to the general characteristics of the cognitive processes that highly g loaded tasks seem to share in common. Little if anything, as yet, is definitely known about the physiological and biochemical substrate of g . What we do already know about g with some assurance, however, is that the measurement of g does not depend on any particular test or types of test or on any particular item contents. These all are merely vehicles, and g can be measured by an incredible variety of vehicles. The elicitation of g does not depend on a specific set of acquired knowledge or skills. As a psychological construct, g cannot be adequately defined in terms of any specific types of information, sets of knowledge, skills, or cognitive strategies. As David Wechsler (1958) has remarked, "... unlike all other factors [g] cannot be associated with any unique or single ability; g is involved in many different types of ability; it is in essence not an ability at all, but a property of the mind" (p.124). Although g is not yet well understood theoretically, it is unquestionably the single largest source of individual differences in all cognitive activities that involve some degree of mental complexity and

that eventuate in some form of behavior which can be measured in terms of some objective standard of performance.

Reference Notes

1. Hennessy, J.J. Structure and patterns of mental abilities in several ethnic groups. Unpublished doctoral dissertation, New York University, 1974. (University Microfilms, No.74-24, 997.)
2. McGurk, F.C.J. Comparison of the Performance of Negro and White High School Seniors on Cultural and Noncultural Psychological Test Questions. Washington, D.C.: Catholic University Press, 1951. (Microcard.)
3. Nichols, P.L. The effects of heredity and environment on intelligence test performance in 4 and 7 year old white and Negro sibling pairs. Unpublished doctoral dissertation, University of Minnesota, 1972.
4. Veroff, J., McClelland, L., & Marquis, K. Measuring intelligence and achievement motivation in surveys. Final Report to U.S Dept. of HEW, OEO, Contract No.OEO-4180. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, October, 1971.

References

- Cattell, R.B. Theory of fluid and crystallized intelligence: A critical experiment. Journal of Educational Psychology, 1963, 54, 1-22.

Cattell, R.B. Abilities: Their Structure, Growth, and Action. Boston: Houghton Mifflin, 1971.

French, J.W., Ekstrom, R.B., & Price, L.A. Kit of Reference Tests for Cognitive Factors. Princeton, N.J.: Educational Testing Service, 1963.

Guilford, J.P. Psychometric Methods. New York: McGraw-Hill, 1954.

Hennessey, J.J., & Merrifield, P.R. A comparison of the factor structures of mental abilities in four ethnic groups. Journal of Educational Psychology, 1976, 68, 754-759.

Jensen, A.R. Educability and Group Differences. London: Methuen (New York: Harper & Row), 1973. a

Jensen, A.R. Level I and Level II abilities in three ethnic groups. American Educational Research Journal, 1973, 4, 263-276. b

Jensen, A.R. Interaction of Level I and Level II abilities with race and socioeconomic status. Journal of Educational Psychology, 1974, 66, 99-111. a

Jensen, A.H. How biased are culture-loaded tests? Genetic Psychology Monographs, 1974, 90, 185-244. b

Jensen, A.R. An examination of culture bias in the Wonderlic Personnel Test. Intelligence, 1977, 1, 51-64.

Loehlin, J.C., Lindzey, G., & Spuhler, J.N. Race Differences in Intelligence. San Francisco: W.H. Freeman, 1975.

McGurk, F.C.J. On white and Negro test performance and socioeconomic factors. Journal of Abnormal and Social Psychology, 1953, 48, 448-450. a

McGurk, F.C.J. Socioeconomic status and culturally-weighted test scores of Negro subjects. Journal of Applied Psychology, 1953, 37, 276-277. b

McGurk, F.C.J. Race differences--twenty years later. Homo, 1975, 26, 219-239.

Paulson, E. An approximate normalization of the analysis of variance distribution. Annals of Mathematical Statistics, 1942, 13, 233-235.

Shuey, A.M. The Testing of Negro Intelligence. (2nd ed.) New York: Social Science Press, 1966.

Spearman, C. The Abilities of Man. New York: Macmillan, 1927.

United States Department of Labor: Manpower Administration. Manual for the USES General Aptitude Test Battery. Washington, D.C.: U.S. Employment Service, 1970.

Wechsler, D. The Measurement and Appraisal of Adult Intelligence. 4th ed. Baltimore: Williams & Wilkins, 1958.

Appendix

Mean white-black differences (in σ units) were determined from data on 24 occupational samples provided in the following Technical Reports of the United States Employment Service of the U.S. Department of Labor Manpower Administration.

Report No.	Date	Occupational Title
S-447	1969	Production-line welder
S-131R74	1974	Fork-lift truck operator
S-180R74	1974	Key-punch operator
S-239R74	1974	Medical ward clerk
S-266R74	1974	Drafter
S-310R74	1974	Electronics Assembler
S-329R74	1974	General office clerk
S-398R74	1974	Teacher Aid (Elementary School)
S-217R75	1975	Banking proof-machine operator
S-228R75	1975	Injection-molding machine tender
S-259R75	1975	Bank Teller
S-270R75	1975	Practical Nurse
S-282R75	1975	Nurse aid
S-370R75	1975	Production and Maintenance Mechanic
S-115R76	1976	Weaver (carpet & rug, textile)
S-135R76	1976	Production-machine operator
S-144R76	1976	Woodworking-machine operator
S-145R76	1976	Grocery checker
S-335R76	1976	Extruding-machine (insulated wire) operator
S-381R76	1976	Electronics micro-logic assembler
S-309R77	1977	Banking encoder
S-414R77	1977	Electrical equipment assembler
S-74R77	1977	Telephone repairer
S-468R78	1978	Cigarette inspector